I CONFERENCIA INTERNACIONAL DE PROCESAMIENTO DE INFORMACION
III Taller de Ciencias de la Información

Agrupamiento de artículos científicos basado en la extracción de n-grams relevantes.

Scientific papers´ clustering based on the extraction of relevant n-grams

Lisvandy Amador Penichet[1], María M. García Lorenzo[2]

1-Lisvandy Amador Penichet. Instituto de Biotecnología de las Plantas, Universidad Central "Marta Abreu" de Las Villas, Cuba. lisvandy@ibp.co.cu
2- María M. García Lorenzo. Departamento de Computación, Universidad Central "Marta Abreu" de Las Villas, Cuba. mmgarcia@uclv.edu.cu

Resumen:

Problemática: La gestión del conocimiento a partir de la información recogida en la bibliografía científica resulta imprescindible para los investigadores en función de optimizar el tiempo de que disponen. El agrupamiento automático de datos se perfila como una de las técnicas que facilitan este proceso. Este permite formar grupos de documentos afines a partir de una colección obtenida mediante un proceso de recuperación de información.

Objetivo(s): Este trabajo tuvo como objetivo: desarrollar un método de agrupamiento para artículos científicos basado en la extracción de n-grams relevantes en las partes del artículo que aportan términos representativos.

Metodología: Los n-grams fueron extraídos de los títulos y de las referencias de los artículos dados que estas subunidades contienen términos que son representativos de los temas que tratan los artículos. Para formar los grupos se tomaron como centroides los n-grams contenidos en la intersección de los títulos de los artículos con los títulos de las referencias. Para evaluar los resultados del método propuesto se utilizaron las medidas externas Overall F-measure, Micro Purity y Macro Purity. Como casos de estudios fueron usados artículos científicos pertenecientes a tres repositorios diferentes.

Resultados y discusión: Se obtuvieron valores por encima de 0.9 para las tres medidas aplicadas en once de los doce corpus sometidos a evaluación. Los valores para el corpus que no superó este umbral estuvieron por encima de 0.7, los cuales se consideran aceptables.

Conclusiones: Los resultados obtenidos a través del experimento realizado demostraron que el agrupamiento de artículos científicos basado en n-grams relevantes se presenta como una alternativa eficaz para la gestión del conocimiento científico.

Abstract:

Problems: *The knowledge´s management based on the information gathered in the scientific literature is essential for researchers in order to optimize the time they have. The data automatic clustering is outlined as one of the techniques that facilitate this process. This allows obtaining related document´ groups formed from a collection obtained through an information retrieval process.*

Objective: *This work aimed to: develop a clustering method for scientific articles based on relevant n-grams' extraction in those parts of the article that provide representative terms.*

Methodology: *n-grams were extracted from the titles and references of the articles given that these subunits contain terms that are representative of the subjects dealing with the articles. To form the groups the n-grams contained in the interception of the articles' titles with the references' titles were taken as centroids. In order to evaluate the results of the proposed method Overall F-measure, Micro Purity and Macro Purity external measures were used. As case studies were used scientific articles belonging to three different repositories.*

Results and discussion: *Values above 0.9 were obtained for the three measures applied in eleven of the twelve corpus submitted to evaluation. The values for the corpus that did not exceed this threshold were above 0.7, which are considered acceptable.*

Conclusions: *The results obtained through the experiment demonstrated that the scientific articles' clustering based on relevant n-grams is presented as an effective alternative as a method of scientific knowledge management.*

Palabras Clave: Agrupamiento de literatura científica; Palabras relevantes; N-grams

Keywords: *Scientific Papers' Clustering; Relevant words; N-grams*

1. Introducción

El desarrollo alcanzado por la ciencia en los últimos años ha traído consigo el incremento constante del número de publicaciones científicas disponibles en internet. Los científicos son los principales consumidores de esta información y generalmente disponen de poco tiempo para procesarla. Por este motivo, se hace necesario desarrollar técnicas y herramientas que permitan gestionar eficientemente el conocimiento científico.

El agrupamiento es uno de los principales métodos usados para la gestión del conocimiento (Qian & Zhang, 2003). El problema del agrupamiento se enfoca en encontrar grupos de objetos similares en un conjunto de datos (Guerrero, Salcedo, & Lamos, 2013). Se desea que los elementos dentro de un mismo grupo presente valores de similitud altos y los que pertenezcan a grupos diferentes presenten valores bajos.

Generalmente el agrupamiento de documentos se basa en el cálculo de una función de similitud que permite determinar el grado de semejanza entre un par de documentos dado. Algunas de las funciones que se han usado con este propósito y han arrojado buenos resultado son la Dice (Vargas Flores, Ledeneva, García Hernández, & Sidorov, 2016), la Coseno y el índice Jaccard (Lin, Jiang, & Lee, 2014).

En (Magdaleno, Fuentes, & García, 2015) se presenta OverallSimSUX que permite agrupar artículos científicos en formato XML teniendo en cuenta la estructura y el contenido de los mismos. Por su parte SimRefBib (Magdaleno, 2015) facilita el agrupamiento de artículos científicos a partir de obtener grupos de documentos más similares tomando en cuenta las referencias bibliográficas.

Al aplicar una de estas funciones a una colección de documentos se obtiene una matriz de similitud donde el elemento en la posición (i,j) de la matriz contiene el grado de semejanza que existe entre este par de documentos. Luego, como resultado de aplicar algún algoritmo de agrupamiento sobre esta matriz, se obtienen los grupos de documentos afines, pudiéndose inferir conocimiento a partir de los grupos formados.

El principal problema asociado a esta forma de agrupamiento es la complejidad computacional que presenta. En los métodos de agrupamiento que siguen esta tendencia es necesario primero procesar los documentos para obtener la representación de los mismos a través de los términos que más se repiten. Luego aplicar alguna función para obtener la matriz de similitud que relaciona los documentos mediante un valor real. Como último paso aplicar un algoritmo sobre esta matriz para

obtener los grupos finales. Los tres pasos tienen una complejidad computacional asociada que generalmente no es baja.

Por otra parte, resulta muy usado para el agrupamiento de artículos científicos considerar el índice de co-citación. De esa forma se determina la similitud entre un par de artículos basado en la cantidad de referencias en común que estos contienen. Algunos de los trabajos desarrollados que hacen uso de índice de co-citación son (Boyack, Small, & Klavans, 2013; Hu, Hu, Gao, & Zhang, 2010; Small, 1973; Small & Sweeney, 1985). La limitante fundamental de los métodos basados en el índice de co-citación es que necesitan disponer de algún índice que recoja las referencias comunes de los artículos que se desea agrupar. Por lo cual, para una colección de artículos de la cual no se disponga las referencias comunes no se puede aplicar un método de este tipo.
Dadas las dificultades anteriormente mencionadas se necesita entonces disponer de métodos de agrupamiento para artículos científicos que eviten estos problemas y se analizan otros aspectos importantes referidos a artículos científicos como es el caso de los n-grams más relevantes de los mismos (Sidorov, Velasquez, Stamatatos, Gelbukh, & Chanona-Hernández, 2014).

Es por ello que el objetivo de este trabajo es desarrollar un método para el agrupamiento de artículos científicos basado en la extracción de n-grams relevantes en las partes del artículo que aportan términos representativos.

La propuesta de método pretende disminuir la complejidad en la extracción de los términos al no tener que procesar todo el artículo., además de eliminar la complejidad asociada al cálculo de la matriz de similitud. Al no trabajarse la co-citación de los artículos las limitantes asociadas a este tipo de agrupamiento quedan totalmente descartadas.

2. Metodología

2.1 Partes del artículo a considerar para la extracción de términos.

El agrupamiento basado en palabras relevantes se fundamenta en la selección de términos que sean representativos de los temas que tratan los artículos de la colección que se desea agrupar.

Un artículo científico puede ser divido en siete partes fundamentales o unidades estructurales: título, autores, afiliación, resumen, palabras claves, contenido y referencias bibliográficas. Estas partes no tienen igual importancia a la hora de seleccionar términos que puedan englobar temas específicos.

Por ejemplo, un autor no puede ser asociado unívocamente a un tema determinado ya que en un periodo de tiempo puede estar vinculado a una línea de investigación y luego vincularse a otra, por lo

cual sus publicaciones no tributarían a un único tema, es por ello que no se tienen en cuenta los autores en el método de agrupamiento que se propone.

La afiliación de los autores queda totalmente descartada para la comparación de dos artículos, ya que dos artículos donde sus autores pertenezcan a una misma institución no tienen por qué tratar un mismo tema. Al mismo tiempo se tiene que dos artículos donde sus autores pertenezcan a instituciones diferentes pueden estar vinculados a una misma temática.

El contenido del artículo sí contiene términos que pueden determinar temas específicos. Sin embargo, la gran cantidad de términos que posee esta unidad estructural provoca que se obtengan tanto términos representativos como términos que no lo son. Estos últimos evidentemente generan ruido a la hora de la comparación de los artículos. Además la gran cantidad de palabras que contiene esta unidad estructural hace que la extracción de términos sea computacionalmente compleja. Por lo cual, al comparar los beneficios de extraer términos de la unidad estructural "contenido" con las desventajas que esta extracción presenta se considera que la complejidad que supone la extracción es mayor que los beneficios que aporta.

En el caso del resumen pasa lo contrario a la sección "contenido". El resumen es un texto que generalmente no debe superar las 250 palabras y debe al mismo tiempo dar una panorámica general de todas las partes del artículos, es decir, debe dar una breve introducción, enunciar de manera concisa el trabajo realizado y mencionar los resultados obtenidos. Debido a esto, la extracción de términos relevantes en esta unidad se ve limitada por la gran variabilidad presentada.

Después de este análisis previo quedan como unidades potenciales para la extracción de términos relevantes: el título, las palabras claves y las referencias bibliográficas.

Las palabras claves constituyen términos que están estrechamente relacionados con el tema que trata el artículo científico. No obstante, estas tienen la restricción que puede que no aparezcan en el título. En muchos casos las palabras que definen los autores como palabras claves, están relacionadas con la temática general que aborda el artículo pero no son las más representativas del tema, ya que las palabras más representativas normalmente forman parte del título.

El título es una única oración, por lo que es lógico pensar que la extracción de términos en esta unidad estructural presenta el mismo problema que el resumen, o incluso peor, ya que cuenta con una cantidad de términos mucho menor. Esto es totalmente correcto, sin embargo, si se analizan todos los

títulos de la colección a la cual se le desea realizar el proceso de agrupamiento como un conjunto, o lo que es lo mismo viendo todos los títulos como un único párrafo y en este párrafo se hace extracción de términos, entonces se obtendrán los términos más relevantes para la colección de manera general. Es por ello que el título del artículo es considerado en la extracción de términos.

Las referencias bibliográficas son consideradas muy importantes a la hora de determinar semejanza entre artículos científicos. Está claro que si dos artículos *i, j* referencian a varios artículos en común, estos dos artículos deben tratar temas similares. Sin embargo, no es necesario hacer uso de toda la información recogida en las referencias cuando se quiere obtener grupos de artículos similares basado en la extracción de términos relevantes. Por ejemplo, el año del artículo citado, las páginas, el volumen y otras informaciones recogidas en las referencias no se consideran relevantes, por esta razón, el método propuesto solo se centra en los títulos de los artículos referenciados. Para la selección de los términos relevantes se toman todos los títulos de un artículo como si fuera un único párrafo y se extraen los términos de este párrafo. Generalmente los títulos de los trabajos contienen palabras que son representativas del tema que el artículo trata, es por ello que la extracción de términos en los títulos de las referencias bibliográficas garantiza que se obtengan palabras representativas del tema en cuestión que trata el artículo.

2.1.1 Método para la extracción de términos relevantes

La extracción de términos relevantes tanto de las referencias bibliográficas como de los títulos de los artículos se realiza siguiendo el procedimiento descrito en (Magdaleno, 2015). El uso de este procedimiento permite la representación de los artículos científicos mediante frases relevantes seleccionadas a partir de los títulos de las referencias bibliográficas, lo cual significa que un documento se representa por los términos que más se repiten, sus respectivas frecuencias de aparición, así como la relación que existe entre estos términos más representativos y la frecuencia de aparición de ellos tomados como frases. Esto evita los problemas asociados a la representación de un documento mediante el Modelo Espacio Vectorial (VSM, por sus siglas en inglés) clásico., uno de los cuales es la "falsa" similitud que se puede encontrar entre un par de documentos que contengan uno o varios términos relevantes en común, pero que estos términos semánticamente tenga significados diferentes.

El siguiente ejemplo ilustra lo anterior. Se tienen dos documentos, uno trata sobre conjuntos difusos (fuzzy set) y otro trata sobre teoría de conjuntos (set theory), en un proceso de recuperación de información, dada una consulta que contenga el término *set* ambos documentos pueden ser

recuperados. Si estos documentos son representados mediante un modelo VSM clásico, en el momento de hallar la similitud entre ellos, el termino *set* aporta cierto grado de similitud que en realidad no existe porque estos documentos tratan temas diferentes.

2.1.2 Método de agrupamiento

Luego del proceso de extracción de términos se tiene:

- el conjunto de términos que se consideran relevantes para la colección, estos son los que se extrajeron de los títulos de los artículos.
- Los términos relevantes para cada artículo. Estos son los que se extrajeron de las referencias bibliográficas de cada artículo.

La idea general del método propuesto consiste en seleccionar aquellos términos de la colección que son representativos. Esto permite obtener una partición del conjunto de documentos donde dos artículos que contengan uno de estos términos pertenecen a un mismo grupo. Los grupos por tanto son determinados por los términos representativos y en primera instancia se tendrán tantos grupos como términos representativos se tengan. A continuación, se analiza cuáles de estos grupos pueden ser unidos ya que se pueden tener varios términos representativos para un tema dado, esto significa que para cada uno de estos términos se obtiene un grupo de documentos, cuando en realidad si dos términos pertenecen a un mismo tema los documentos asociados a estos términos deben pertenecer a un mismo grupo.

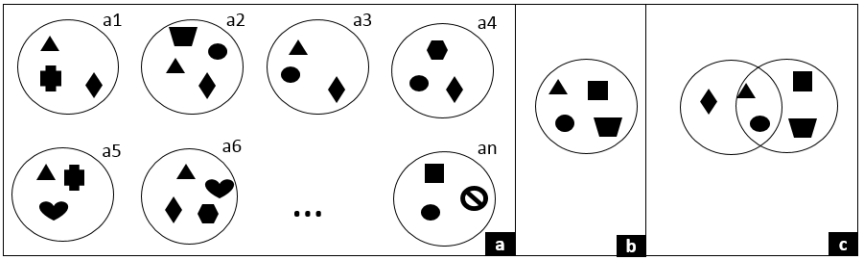

Figura 1. Selección de términos centroides. a) palabras relevantes para cada documento seccionadas a partir de los títulos de las referencias. b) palabras relevantes de la colección seleccionadas a partir de los títulos de los artículos. c) intersección de las palabras relevantes de los títulos con las palabras relevantes de las referencias.

Para obtener los términos que son finalmente usados como centroides para la conformación de los grupos, se intersectan los términos que se obtuvieron a partir de los títulos de los artículos pertenecientes a la colección (Figura 1b), con los términos relevantes que se obtienen a partir de los títulos de las referencias bibliográficas de todos los artículos (Figura 1a). Los términos relevantes de

las referencias bibliográficas son aquellos que más se repiten entre los términos seleccionados como relevantes para cada artículo.

En la Figura 1c se aprecia que la intersección de las palabras relevantes de las referencias con las palabras relevantes de los títulos permite obtener finalmente los términos que serán los centroides de los grupos a formar. Para formar los grupos preliminares se une a cada uno de estos centroides los documentos que los contengan. Esto indica que en este paso puede haber documentos que se agreguen a más de un grupo (Figura 2).

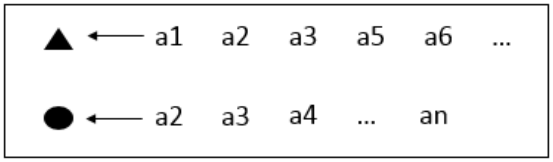

Figura 2. Ubicación inicial de los artículos en los grupos correspondientes basada en las palabras relevantes que estos contienen.

Para eliminar el solapamiento de los grupos se construye un grafo de conexiones de los documentos. Dos documentos están conectados si contienen algún término relevante en común. Luego se extraen de los grupos los elementos que provocan solapamiento y se analiza la pertenencia promedio de estos elementos a cada uno de los grupos que solapan. La pertenencia promedio de un elemento *i* a un grupo C_j se calcula como sigue:

$$PertProm(i, C_j) = \frac{\sum_{k=1}^{l} M(i, C_{jk})}{l} \quad (1)$$

Donde l representa la cantidad de elementos del grupo C_j y M es la matriz de adyacencia que representa el grafo de conexiones entre los artículos.

Entrada: Conjunto de *n* artículos científicos $(A = \{a_1, a_2, \dots, a_n\})$
Salida: Lista de clústers formados (C)
Inicio:
P1. Conformar Lista *PRT* con palabras relevantes de los títulos.
P2. Para cada artículo a_i conformar Lista $PRRA_i$. $PRRA_i$ palabras relevantes de las referencias del artículo i
P3. Conformar Lista PRR con los términos relevantes de $\bigcup_{i=1}^{n} PRRA_i$
P4. Lista de palabras centroides PC=$PRT \cap PRR$
P5. Para cada a_i Si $PC_j \subseteq PRRA_i$ asignar a_i a C_j, donde PC_j es el centroide de C_j
P6. Si $cup \leftarrow \bigcup_{g,h=1}^{k} cap(a_g, a_h) = \emptyset$, donde $cap = a_g \cap a_h$, entonces ir al paso 8.
P7. Determinar para cada $a_i \in cap(c_g, c_h) \neq \emptyset$, el C_j correspondiente, donde $PertProm$ es máxima. Ir paso 6. $PertProm$: pertenencia promedio
P8. $\forall a_i \notin C, C_j \leftarrow (C_j \cup a_i)$, donde $PertProm$ es máxima.
Fin

Figura 3. Algoritmo de agrupamiento basado en términos relevantes.

El elemento es finalmente añadido al grupo con respecto al cual tiene mayor pertenencia promedio. De este mismo modo se procede con los elementos que en primera instancia no fueron unidos a ningún grupo. Esto ocurre debido a que dichos documentos no contienen dentro de sus términos representativos ninguno de los términos que fueron seleccionados como centroides. Debido a esto se requiere calcular la pertenencia promedio de estos elementos a cada uno de los grupos. En la Figura 3 se formaliza el algoritmo de agrupamiento basado en la extracción de términos relevantes.

3. Resultados y discusión

Para evaluar la eficacia del método propuesto se diseñó un experimento basado en la aplicación de medidas externas. Las medidas seleccionadas fueron la medida *Overall F-measure* (OFM) propuesta en (Steinbach, Karypis, & Kumar, 2000) y las medidas *Micro Purity* y *Macro Purity* propuesta su utilización por INEX(Costa & Ortale, 2013). Estas tres medidas se basan en una estructura previamente especificada que refleja la intuición que se tiene del agrupamiento de los datos. La elección de estas se sustenta en que permiten describir la calidad del resultado completo del agrupamiento usando un único valor real (Amador, García, Gálvez, & Magdaleno, 2017).

Como casos de estudios se utilizaron artículos provenientes del sitio ICT[1], artículos pertenecientes al repositorio IDE-Alliance y artículos de la revista Springer cuya clasificación de referencia se obtuvo mediante la herramienta AuthorMapper[2]. En la

Tabla 1 se especifican las características de cada uno de los corpus utilizados. Note que los cinco últimos corpus fueron creados tomando documentos de los tres repositorios.

La Figura 4 muestra los valores obtenidos para las medidas externas usadas al aplicar el algoritmo propuesto a cada uno de los corpus sometidos a evaluación. Se puede observar que excepto para el último corpus, los valores obtenidos siempre fueron mayores que 0.9. Esto significa que los grupos obtenidos con el método propuesto son bastante similares a los grupos presentados como clasificación de referencia; lo que demuestra que el algoritmo basado en la extracción de n-grams

[1] ftp://ict.cei.uclv.edu.cu
[2] www.authormapper.com

relevantes es capaz de discernir correctamente entre los elementos que son similares y los que no lo son.

Tabla 1. Descripción de los archivos utilizados para evaluar la calidad del agrupamiento.

No. Corpus	Cantidad de documentos	Cantidad de clases	Temas que trata
Conjuntos de documentos XML confeccionados a partir de documentos recuperados del sitio del ICT del Centro de Investigaciones en Informática de la Universidad Central "Marta Abreu" de Las Villas http://ict.cei.uclv.edu.cu			
1	32	2	Fuzzy Logic (FL), SVM
2	25	2	Rough Set (RS), Association Rules
3	32	2	RS, SVM
4	28	2	Association Rules (AR), Fuzzy Logic
5	32	2	Association Rules, SVM
Recopilación de documentos del repositorio IDE-Alliance, internacionalmente utilizados para evaluar agrupamiento. Proporcionados por la Universidad de Granada. España.			
6	19	2	Copula, Belief Propagation
Documentos de la revista Springer cuya clasificación de referencia se obtuvo mediante la herramienta AuthorMapper			
7	30	2	Oxidative stress, Gene Expression
Documentos pertenecientes al sitio ICT, al repositorio IDE-Alliance y al sitio			
8	41	4	RS, Copula, Belief Propagation (BP), CL
9	29	2	Copula, SVM
10	38	3	Copula, SVM, BP
11	30	2	FL, Gene Expression
12	89	7	AR, FL, SVM, Copula, BP, CL, RS

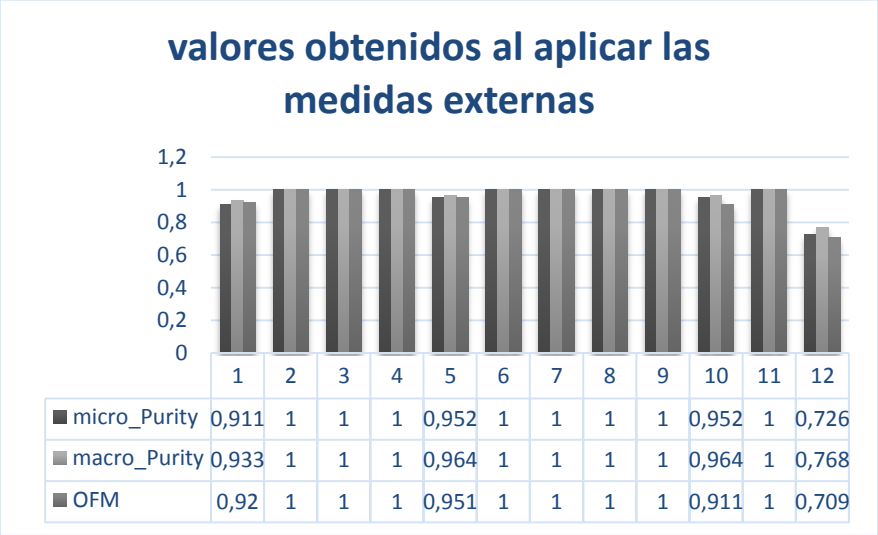

Figura 4. Resultados obtenidos para las medidas, *Micro Purity*, *Macro Purity* y *OFM* al aplicar el algoritmo de agrupamiento basado en palabras relevantes.

La disminución de la calidad del agrupamiento en el último corpus, manifestada en el descenso de los valores obtenidos para las medidas aplicadas, puede estar asociada al elevado número de clases que conforman este corpus. Aun así estos valores se consideran aceptables. Además es válido aclarar que no es común que las colecciones sometidas a procesos de agrupamiento contengan más de dos o tres clases. Lo común es que se tengan dos clases: (1) la clase a la que pertenecen los documentos que tratan el tema específico que es de interés para el usuario, (2) la clase de los documentos que a pesar de ser devueltos en el proceso de recuperación de información no están vinculados con el tema en cuestión que es de interés para el usuario. En última instancia esta segunda clase se puede ver con un conjunto de *k* subclases donde *k* es la cantidad de documentos que la integran. Sin embargo, considerarla de esta manera tampoco afecta el proceso de agrupamiento ya que se tiene un conjunto de documentos aislados, y un conjunto de documentos agrupados los cuales forman la clase que es en efecto de interés para el usuario.

4. Conclusiones

El agrupamiento de artículos científicos constituye una técnica eficiente para la gestión del conocimiento a partir de la bibliografía disponible. Sin embargo, los métodos de agrupamiento con que se cuenta presentan algunas limitaciones que no permiten explotar al máximo las características de estos documentos.

El método de agrupamiento propuesto, basado en la extracción de n-grams relevantes a partir de los títulos y de las referencias bibliográficas, permite obtener grupos de documentos que satisfagan las

necesidades de información de los usuarios. Este método reduce la complejidad computacional ya que no hace selección de términos en todo el documento. Además no presenta las limitantes de los métodos de agrupamiento de artículos científicos basados en los índices de co-citación, debido a que no se basa en las referencias comunes para hallar similitud entre los documentos.

5. Referencias bibliográficas

1 Amador, L., García, M. M., Gálvez, D., & Magdaleno, D. (2017). SemClustDML: algoritmo para agrupar artículos científicos basado en la información brindada por las referencias bibliográficas. *Revista Cubana de Ciencias Informáticas, 11*(2), 46-60.

2 Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology, 64*(9), 1759-1767.

3 Costa, G., & Ortale, R. (2013). *A latent semantic approach to xml clustering by content and structure based on non-negative matrix factorization.* Paper presented at the Machine Learning and Applications (ICMLA), 2013 12th International Conference on.

4 Guerrero, C. D., Salcedo, D., & Lamos, H. (2013). A Clustering Approach to Reduce the Available Bandwidth Estimation Error. *IEEE Latin America Transactions, 11*(3), 927-932.

5 Hu, C.-P., Hu, J.-M., Gao, Y., & Zhang, Y.-K. (2010). A journal co-citation analysis of library and information science in China. *Scientometrics, 86*(3), 657-670.

6 Lin, Y.-S., Jiang, J.-Y., & Lee, S.-J. (2014). A similarity measure for text classification and clustering. *IEEE transactions on knowledge and data engineering, 26*(7), 1575-1590.

7 Magdaleno, D. (2015). *Metodología para el agrupamiento de documentos semiestructurados.* (Dr.), Universidad Central "Marta Abreu" de Las Villas, Santa Clara.

8 Magdaleno, D., Fuentes, I. E., & García, M. M. (2015). Clustering XML Documents Using Structure and Content based on a New Similarity Function OverallSimSUX. *Computación y Sistemas, 19*(1), 151-161.

9 Qian, Y., & Zhang, K. (2003). *A customizable hybrid approach to data clustering.* Paper presented at the Proceedings of the 2003 ACM symposium on Applied computing, New York.

10 Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications, 41*(3), 853-860.

11 Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the Association for Information Science and Technology, 24*(4), 265-269.

12 Small, H., & Sweeney, E. (1985). Clustering the science citation index® using co-citations: I. A comparison of methods. *Scientometrics, 7*(3-6), 391-409.

13 Steinbach, M., Karypis, G., & Kumar, V. (2000). *A comparison of document clustering techniques.* Paper presented at the KDD workshop on text mining.

14 Vargas Flores, S. I., Ledeneva, Y. N., García Hernández, R. A., & Sidorov, G. (2016). Comparación de medidas de similitud para desambiguación del sentido de las palabras utilizando rankeo de grafos.